3LB-SAT: Una herramienta de anotación semántica

Empar Bisbal, Antonio Molina, Lidia Moreno, Ferran Pla, Maximiliano Saiz-Noeda*, Emilio Sanchis

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
{ebisbal, lmoreno, amolina, fpla, esanchis} @dsic.upv.es

*Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
max@dlsi.ua.es

Resumen: Presentamos una herramienta, llamada 3LB-SAT, para el etiquetado semántico de corpus multilingüe. Sus principales características son que está orientado a la palabra, que permite introducir el corpus en diferentes formatos (formato parentizado TBF y formato XML) y que usa el diccionario *EuroWordnet* para consultar el sentido de las palabras en cuatro lenguas (español, catalán, euskara e inglés).
Palabras clave: lenguaje natural, corpora anotados, herramientas de anotación semántica.

Abstract: We present a tool, called 3LB-SAT, for the semantic tagging of multilingual corpora. Main features of this tool are that it is word-oriented, allows different formats for input corpus (TBF format, PenTreebank Bracketted Format and XML) and uses EuroWordnet for searching the word sense in four languages.
Keywords: natural language, corpora annotated, semantic annotation tool.

1 Introducción

Durante los últimos años se han realizado numerosos esfuerzos en el desarrollo de recursos lingüísticos como diccionarios, bases de datos léxicas y corpus anotados con diferente información lingüística.

Estos recursos han propiciado el creciente auge de la lingüística computacional, tanto de las aproximaciones inductivas o basadas en corpus, que utilizan los textos como datos de aprendizaje y evaluación, como de las aproximaciones deductivas que usan estos recursos como base de conocimiento y también de evaluación.

Desde un punto de vista puramente lingüístico, los corpus son bases de datos imprescindibles para el estudio de la lengua ya que proporcionan ejemplos analizados/anotados de lenguaje real. El estudio lingüístico revierte directamente en la mejora de la calidad de los recursos, dotándolos de una mayor robustez.

Desafortunadamente la mayor parte de estos recursos sólo están disponibles para lenguas como el inglés. En ese sentido, podemos encontrar corpus anotados con información morfológica y sintáctica, como el *Penn Treebank II* (Marcus *et al*., 1993) o con información semántica, como el *SemCor* (Miller *et al*., 1993,1994).

A pesar de los grandes esfuerzos dedicados recientemente en el desarrollo de recursos lingüísticos, estas colecciones de textos para lenguas mayoritarias como el castellano, presentan carencias en cuanto al volumen, fiabilidad y accesibilidad. Aunque existen grandes corpora textuales y orales para el castellano, estos generalmente no están anotados con información lingüística[1], o si lo están, como el corpus LexEsp, anotado con información morfosintáctica, sólo una pequeña parte del mismo está supervisada[2].

Este problema todavía se ve agravado en otras lenguas minoritarias del estado español como el catalán y el euskara que aunque disponen de algunos corpus, estos son insuficientes en cuanto a su volumen.

[1] Ver Corpus de Referencia del Español Actual (CREA) en http://corpus.rae.es/creanet.html

[2] El corpus LexEsp pertenece al proyecto del mismo nombre llevado a cabo por el Departamento de Psicología de la Universidad de Oviedo, y desarrollado por el Grupo de Lingüística Computacional de la Universidad de Barcelona, con la colaboración del Grupo de Procesamiento del Lenguaje de la Universidad Politécnica de Cataluña. Véase http://clic.fil.ub.es/demos/corpus/

Con el fin de afrontar estas carencias surge el proyecto "3LB: Construcción de una base de datos de árboles sintáctico semánticos"[3] en el que intervienen cinco grupos de Procesamiento del Lenguaje Natural (PLN) del estado español: Universidad de Alicante, Universidad de Barcelona, Universidad del País Vasco, Fundación Bosch i Gimpera, y Universidad Politécnica de Valencia.

A pesar de que la supervisión de corpora es una tarea costosa, creemos que es una labor imprescindible para el desarrollo de aplicaciones reales en el área del Procesamiento del Lenguaje Natural y, como tal, para la mejora de la calidad de todos los sistemas que requieran PLN como son la Traducción Automática, la Extracción de Información, la Recuperación de Información, el Resumen Automático o la Búsqueda de Respuestas.

Con el fin de facilitar la tarea de supervisión de los distintos niveles de anotación de textos por lingüistas es necesario el uso de herramientas de ayuda a esta tarea.

En este artículo se presenta la descripción y funcionalidades de un conjunto de herramientas para este fin. Primero se hace una breve descripción del proyecto 3LB, a continuación se describen las estrategias y recursos utilizados en cada uno de los niveles lingüísticos considerados, centrándonos principalmente en la herramienta de anotación semántica desarrollada. Por último se presenta una serie de conclusiones y trabajos futuros.

2 Proyecto 3LB

El objetivo principal de este proyecto es construir tres corpus anotados sintácticamente (*treebanks*) para el español, catalán y euskara. Además de la anotación sintáctica, se realizará una anotación semántica mediante los *synsets* de los diferentes Wordnets (FellBaum, 1998) elaborados en cada lengua, así como una anotación de los elementos anafóricos, elípticos y correferentes. Para el español y el catalán el volumen del corpus será de 100.000 palabras cada uno, en el caso del euskara 50.000 por razones de mayor complejidad notacional y menor cobertura del Wordnet de que se dispone (35.000 entradas frente a las 100.000 existentes para el castellano o las 65.000 para el catalán).

[3] Para más información sobre este proyecto, visitar http://www.dlsi.ua.es/projectes/3lb.

El corpus CLiC-TALP[4] para el español y el catalán consta actualmente de 100.000 palabras anotadas morfológica y sintácticamente de manera automática. El etiquetado morfosintáctico está supervisado por lingüistas mientras que la anotación sintáctica se encuentra en periodo de revisión.

El resto del corpus, hasta 5,5 millones de palabras está anotado morfosintácticamente de forma automática, con una tasa de error estimado del orden de un 3 %.

El corpus del que disponemos para el euskara en este proyecto consta de 40.000 palabras anotadas morfosintácticamente de manera manual. En este proyecto se trata de etiquetar este corpus sintáctica y semánticamente según la propuesta y ampliarlo hasta 50.000 palabras con anotación morfológica, sintáctica y semántica.

Se cuenta con corpora en diferentes formatos: formato parentizado (TBF, Treebank Bracketted Format; puede verse un ejemplo en la Figura 1) y el formato de etiquetado basado en la tecnología XML. Para este último se ha definido en el marco del proyecto un DTD (*Document Type Definition*) que permite describir la estructura de la información almacenada para cada fichero del corpus:

- Información genérica como el idioma, el estado de anotación sintáctica y semántica (en proceso o supervisada) y las versiones de EWN y WN que han sido utilizadas.
- Información sobre el lema asociado a las palabras y su categoría morfosintáctica (POS), así como la estructura sintáctica de las oraciones, definiendo la jerarquía existente entre los distintos constituyentes.
- Información sobre la anotación semántica, para lo que se almacenan los identificadores de los *synsets* del *EuroWordnet* asociados a cada palabra del corpus.

3 Estrategia de anotación

La estrategia de anotación morfológica y sintáctica de los corpora se ha realizado de manera semiautomática intentando facilitar la tarea mediante el uso de herramientas de desambiguación automática (desambiguación morfosintáctica y sintáctica) que proporcionan unos niveles de precisión aceptables para su uso en el

[4] Véase http://clic.fil.ub.es/recursos/corpus.shtml

proceso de supervisión. Además de los niveles sintáctico y semántico, la anotación también trata una tercera dimensión, el etiquetado correferencial en el que se abordarán algunos fenómenos de elipsis y anáfora así como el establecimiento de las cadenas de correferencia.

En el caso de la anotación automática de la semántica, los métodos actuales obtienen unos niveles de precisión que no son lo suficientemente adecuados para su utilización en el proceso de supervisión. Debido a esto, se ha optado por definir una estrategia totalmente manual diseñando una herramienta que, utilizando los recursos lingüísticos disponibles, facilite la tarea a los anotadores. Sin embargo, se ha dejado abierta la posibilidad de utilizar un sistema de desambiguación semántica automático, bien exigiendo que el formato de salida del desambiguador se ajuste al XML definido en el proyecto, o bien incorporándolo a la herramienta.

3.1 Anotación sintáctica

Para realizar la anotación sintáctica se dispone de varias herramientas de análisis sintáctico desarrolladas por los grupos participantes en el proyecto. Todas ellas realizan análisis sintáctico con distinta profundidad: APOLN (Molina *et al*., 1999), es un analizador ascendente basado en expresiones regulares que puede segmentar oraciones del castellano en sintagmas básicos o 'chunks'; TACAT[5] (Atserias *et al*., 1999) es un analizador ascendente que realiza análisis sintáctico parcial tanto para el castellano como para el catalán, para los cuales tiene definida la correspondiente gramática incontextual; y SUPP (Martínez-Barco *et al*., 1998), es un analizador descendente que realiza análisis sintáctico parcial para el castellano identificando también sintagmas recursivos. Ninguno de estos analizadores está evaluado de forma fiable, ya que no se disponen de corpora de referencia supervisados. Por lo tanto, la elección del analizador se basó en otros criterios: el número de lenguas capaz de analizar y un grado de profundidad del análisis adecuado.

En este sentido, se consideró que un análisis sintáctico demasiado profundo (con estructuras recursivas, sintagmas verbales, etc.) no sería muy adecuado porque para ese grado de análisis las prestaciones de los analizadores disminuyen, y esto lleva consigo que durante el proceso de supervisión aparezcan árboles de análisis incorrectos cuya corrección es muy costosa. Es preferible disponer de un análisis parcial muy preciso a partir del cual, sin mucho esfuerzo por parte del supervisor, pueda construirse el árbol sintáctico completo.

Por esos dos motivos se escogió TACAT para el castellano y para el catalán. Respecto al euskara, el corpus disponible es EPEC (Aduriz *et al*., 2002, 2003), y las herramientas utilizadas son EUSLEM (lematizador) y ZATIAK (*chunker*).

Para la supervisión de los textos anotados sintácticamente se ha utilizado el *TreeTrans*, una herramienta de libre disposición perteneciente al grupo de herramientas de anotación gráfica AGTK[6] (Bird *et al*., 2002). Esta herramienta permite la visualización y la supervisión de textos. Además posibilita la anotación morfológica y sintáctica, para lo que emplea una estructura en forma de árbol donde los nodos pueden ser de tres tipos: nodo con información sintáctica (syn), nodo con información morfológica (pos) o nodos hojas que contienen las palabras del texto original (wrd). Pero, una vez más, la carencia de recursos desarrollados en el campo de la anotación semántica supuso un problema: no disponíamos de una herramienta para la supervisión de este tipo de anotación. Por ello, se ha modificado el *TreeTrans* para que permita la supervisión de ambos niveles de anotación y para que acepte los formatos de entrada/salida definidos en el marco del proyecto.

```
(( S
   ( sn
     ( espec.ms
       ( da0ms0 El ))
     ( grup.nom.ms
       ( ncms000 gato )))
   ( grup.verb
     ( vmip3s0 come ))
   ( sn
     ( grup.nom.ms
       ( ncms000 pescado )))
   ( Fp . )))
```

Figura 1. Formato TBF

[5] Desarrollado por el Grupo de Investigación de Lenguaje Natural del Departamento de Lenguajes y Sistemas Informáticos de la Universidad Politécnica de Cataluña en colaboración con el Laboratorio de Lingüística Computacional de la Universidad de Barcelona. Demostración del analizador disponible en http://nipadio.lsi.upc.es/cgi-bin/demo/demo.pl

[6] Véase http://agtk.sourceforge.net/

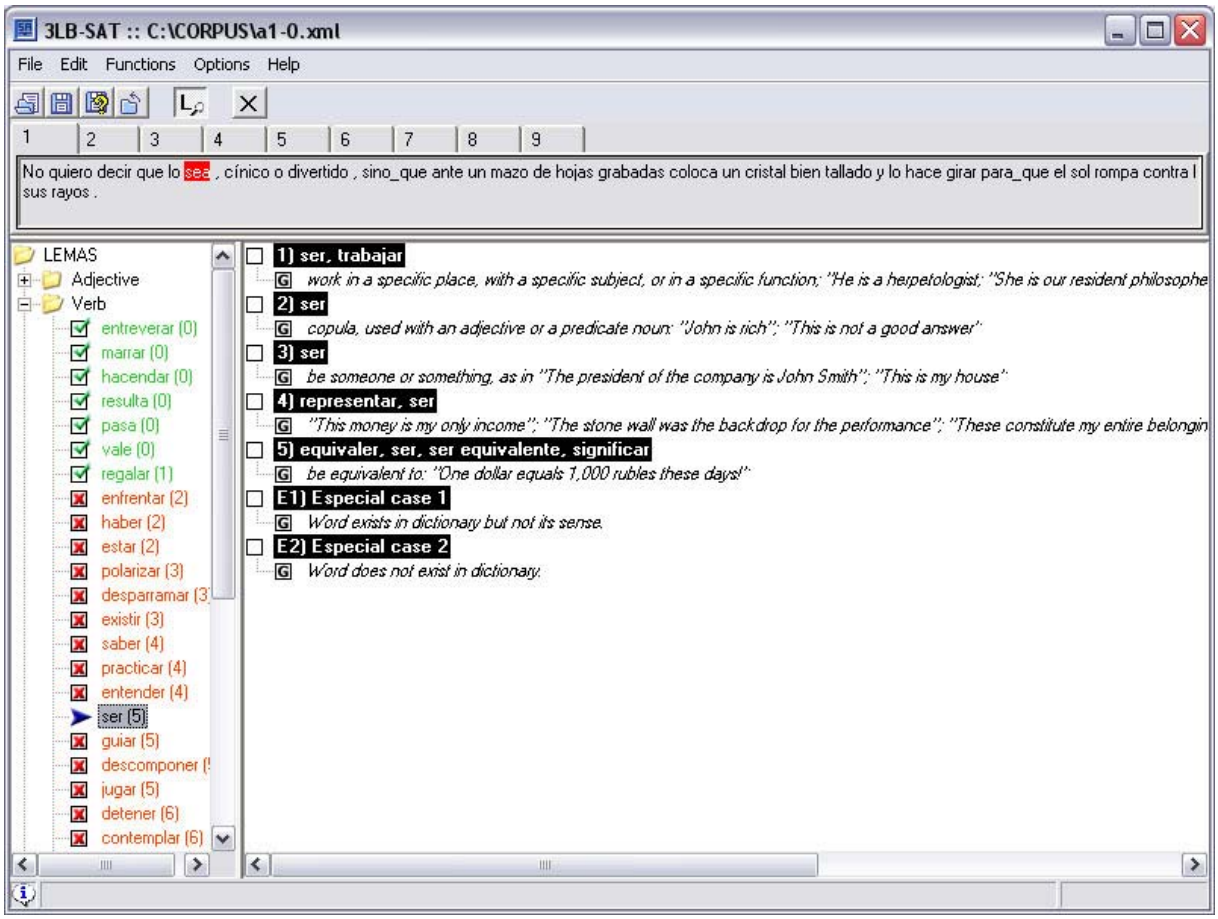

Figura 2. Sentidos para la palabra *ser*

3.2 Anotación semántica

El objetivo de realizar la anotación semántica del corpus de manera paralela a la sintáctica condujo a la modificación de la herramienta *TreeTrans*. Así, el proceso de anotación semántica se realizaría frase a frase, de forma que se anotarían las palabras a medida que aparecen en las frases.

Sin embargo, la anotación semántica no es una tarea trivial ya que una palabra puede tener muchos sentidos diferentes y la diferencia entre ellos es, en ocasiones, muy sutil. Por esta razón, los anotadores humanos deben estudiar concienzudamente todos los sentidos de una palabra antes de ser capaces de anotarla. Se convino, por ello, en la creación de una herramienta paralela orientada a los tokens, palabras o conjuntos de palabras agrupados utilizando un analizador morfológico (para permitir la anotación de locuciones, unidades multipalabra, nombres compuestos, etc.). De esta manera, el corpus no se anotaría de manera secuencial, sino que, dado un lema, el sistema permitiera anotar todas sus apariciones en el corpus (o en diferentes segmentos del mismo). Este recorrido no secuencial mejoraría sustancialmente el tiempo y esfuerzo invertido en la anotación. Posteriormente, un recorrido secuencial del texto anotado semánticamente, frase por frase, mediante la herramienta *TreeTrans* que, como ya hemos mencionado, se ha modificado para que acepte este tipo de anotación, permitiría la supervisión del mismo.

Así es como nace la herramienta 3LB-SAT (3LB-Semantic Annotation Tool). Sus principales características son que está orientado a la palabra (o token), que permite introducir el corpus en diferentes formatos (TBF y XML) y que usa *EuroWordnet* para consultar el sentido de las palabras en cuatro lenguas (español, catalán, euskara e inglés) como puede verse en la Figura 3. Las palabras monosémicas se anotan automáticamente; y se muestran todas las apariciones de un lema en el texto, siendo posible asociar más de un *synset* con una aparición de un lema.

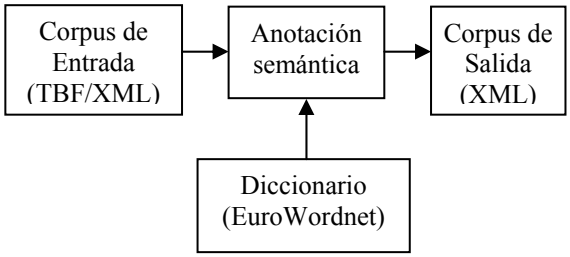

Figura 3. Esquema de la herramienta 3LB-SAT

Como se ha comentado, 3LB-SAT usa *EuroWordnet* para consultar el sentido de las palabras ya que se dispone de éste para las cuatro lenguas objetivo de nuestro proyecto. De forma simplificada, este diccionario consiste en conjuntos de sinónimos (*synsets*) que agrupan los sentidos de distintas palabras asociados a un único concepto. Cada uno de estos *synsets* tiene asociado un identificador único utilizado para anotar los sentidos con la herramienta. En el caso de cambio de identificador entre las distintas versiones de *WordNet*, un mapeo básico entre versiones permite la adaptación del etiquetado.

Se han identificado dos posibles tipos de carencias de *EuroWordnet* que se desea que también sean anotadas: (a) La no existencia del sentido que representa la palabra dentro de la oración (etiquetado como *C1S*); (b) La no existencia de la palabra en *EuroWordnet* (etiquetado como *C2S*). Esta anotación especial nos permitirá enriquecer el diccionario con nuevos sentidos para las palabras existentes o ampliarlo a nuevas palabras.

La herramienta debe conocer el idioma del corpus para buscar en el *EuroWordnet* correspondiente; si la entrada está en formato XML, la herramienta lo lee de la cabecera del fichero, pero si está en formato TBF lo debe indicar el usuario.

Durante la apertura de un fichero se anotan de forma automática todas las palabras monosémicas (aunque éstas han de ser supervisadas por si este sentido no fuera apropiado, caso que debería anotarse con la etiqueta C1S) y aquellas que no se han encontrado en el *EuroWordnet*, a las que se les asigna la etiqueta C2S como se ha comentado con anterioridad.

Una vez cargado el corpus, la herramienta muestra en la parte izquierda todos los lemas siguiendo un código de colores para distinguir que no se ha anotado ninguna aparición del lema en el corpus, o que se ha anotado alguna de sus apariciones, pero no todas, o bien todas sus apariciones han sido anotadas. Además, los lemas se muestran por categorías en el orden seleccionado: ascendente/descendente por orden polisémico o por orden alfabético.

Cuando se selecciona un lema, se muestran todas sus apariciones en la parte superior de la ventana. Si se selecciona una de ellas, se muestran todos los posibles sentidos del lema (para cada *synset* se muestran todos los sinónimos y la glosa); si no hay una glosa asociada al *synset* la herramienta buscará la glosa del *synset* equivalente en el *Wordnet* inglés. Una vez se ha seleccionado el lema y una de sus apariciones se anota(n) su(s) sentido(s). En la Figura 2 podemos ver como el lema *ser* aparece nueve veces en el texto en la categoría verbal, y tiene cinco sentidos.

Durante el proceso de anotación, la herramienta creará un informe relativo a los cambios efectuados sobre un fichero del corpus durante una sesión de trabajo. Esta información nos permitirá obtener estadísticas, comparar el proceso de anotación utilizando métodos automáticos de desambiguación o sin ellos, además de poder realizar un seguimiento del sistema.

3.3 Anotación correferencial

Otra de las dimensiones de etiquetado enmarcadas en el proyecto es la anotación correferencial. Este tipo de anotación es prácticamente inexistente en los recursos disponibles tanto comercial como gratuitamente y su desarrollo, tanto para el español, como para el resto de las lenguas involucradas en el proyecto, es fundamental para la evaluación de sistemas de resolución de este tipo de fenómenos lingüísticos.

Esta anotación afecta a fenómenos de relevancia como la elipsis y la anáfora. Así mismo, la fase de anotación pretende establecer cadenas de correferencia que relacionen entre sí a todos los elementos textuales que compartan referente en el corpus.

Para la anotación anafórica, se hará uso de un sistema de etiquetado asistido, esto es, se utilizará un sistema automático de resolución de la anáfora que se encargará de detectar y proponer el antecedente correcto de un elemento anafórico. Esta información será transferida al anotador humano que se encargará de verificar su corrección y aceptar o modificar la solución propuesta para completar la fase de etiquetado. Para ello, contará con una interfaz gráfica pro-

vista de un conjunto de códigos de colores para identificar con mayor facilidad todos los elementos que intervienen en el etiquetado.

Si bien el fenómeno de la anáfora es extremadamente amplio y complejo, con el fin de plantear una anotación anafórica abordable, se han incluido dentro de los objetivos marcados en este proyecto el etiquetado anafórico tanto de las elipsis de sujeto y adjetivas como de las anáforas pronominales, dentro de las cuales se incluyen los pronombres átonos de sujeto y los de complemento dentro de un sintagma preposicional así como los pronombre clíticos.

Con la incorporación de este etiquetado correferencial, el corpus adquirirá una dimensión de gran interés, especialmente para la evaluación de sistemas de resolución de la anáfora.

4 Conclusiones y direcciones futuras

En este artículo hemos presentado una herramienta de anotación semántica que cuenta con las características deseadas para este tipo de sistemas: multiplataforma, multilingüe y de fácil uso.

Por ello, 3LB-SAT se ha implementado utilizando el lenguaje de programación *python*, lenguaje que proporciona a la herramienta la versatilidad y la portabilidad entre distintas plataformas como Windows o Linux.

Otras funcionalidades están todavía en desarrollo, como la propagación automática de la etiqueta semántica, es decir, cuando se anota la primera aparición de una palabra esta etiqueta se propone para el resto de apariciones de la palabra en el corpus.

Esta herramienta facilita la obtención de corpora etiquetados con información semántica. Además, los corpora resultantes permitirán evaluar, mejorar y enriquecer los recursos lingüísticos, y los sistemas automáticos de análisis morfológico, sintáctico y semántico sirviendo como base de desarrollo de herramientas lingüísticas en diferentes ámbitos de aplicación como la traducción automática, el acceso en lenguaje natural (castellano, catalán, eusquera o inglés) a sistemas de información como la web, o las bases de datos documentales.

Actualmente la herramienta está siendo probada por los anotadores para detectar posibles problemas o dificultades de uso. Ya han sido establecidos los criterios de anotación y está previsto que en un plazo de cuatro meses se podrán tener anotados gran parte de los corpora utilizados en el proyecto 3LB.

Se está trabajando también en la incorporaciópn a la herramienta de una serie de utilidades que permitan obtener información estadística de la efectividad de la tarea de anotación semántica, índice de acuerdo entre los diferentes anotadores, etc., con el fin de determinar la estrategia de anotación más eficiente.

5 Agradecimientos

Este trabajo ha sido subvencionado por el proyecto PROFIT 3LB (FIT-150500-2002-244) y el proyecto CICYT TUSIR (TIC2000-0664-C02-01)
Agradecemos a los miembros de los grupos participantes en el proyecto “3LB: Construcción de una base de datos de árboles sintáctico semánticas” su ayuda para la confección de este artículo.

Bibliografía

Aduriz I., I. Aldezabal, M. Aranzabe, B. Arrieta , J. Arriola, A. Atutxa, A. Díaz de Ilarraza, K. Gojenola, M. Oronoz y K. Sarasola (2002). Construcción de un corpus etiquetado sintácticamente para el euskera. Actas del XVIII Congreso de la SEPLN Universidad de Valladolid. Valladolid. España.

Aduriz I., M. Aranzabe, J. Arriola, A. Atutxa, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa y R. Urizar (2003). *Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing* Corpus Linguistics Around the World. Language and Computers. Ed. Andrew Wilson, Paul Rayson and Dawn Archer. Rodopi. Netherlands. (en prensa)

Atserias, J., J. Carmona, I. Castellón, S. Cervell, M. Civit, L. Màrquez, M. A. Martí, L. Padró, R Placer, H. Rodríguez, M. Taulé y J. Turmo" (1998). *Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text*, Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC, pp. 1267-1272, Granada, Spain.

Bird, S., K. Maeda, X. Ma, H. Lee, B. Randall, y S. Zayat (2002). *TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse Tools*

Built on the Annotation Graph Toolkit. In Proceedings of the Third International Conference on Language Resources and Evaluation, Paris.

Fellbaum, C. (1998). *WordNet, an electronic lexical database*. MIT Press.

Ferrández, A., M. Palomar y L. Moreno, (1998). *Anaphor resolution in unrestricted texts with partial parsing*, Proceedings of COLING-ACL'98, Montreal, Canadá.

Mahesh, K. and S. Nirenburg. (1995a). *A Situated Ontology for Practical NLP*. In Proc. Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95), Aug. 19-20, 1995. Montreal, Canada.

Mahesh, K. y S. Nirenburg. (1995b). *Semantic Classification for Practical Natural Language Processing*. In Proc. Sixth ASIS SIG/CR Classification Research Workshop: An Interdisciplinary Meeting. October 8, 1995, Chicago IL.

Marcus, M.P.; B. Santorini y M.A. Marcinkiewicz (1993). *Building a large annotated corpus of English: the Penn treebank.* Computational Linguistics, 19(2), pp.313-330.

Martínez-Barco, P., J. Peral, A. Ferrández, L. Moreno y M. Palomar (1998). Analizador Parcial SUPP. Proceedings of the 6th Iberoamerican Conference on Artificial Intelligence (IBERAMIA'98), pp 329--341", Lisbon, Portugal.

Mezquita, Y. L., G. Sidorov y A. Felbukh (2003). *Tool for Computer-Aided Spanish Word Sense Disambiguation*. In Computational Linguistics and Intelligent Text processing, Springer-Verlag.

Miller G. A., M. Chodorow, S. Landes, C. Leacock y R. G. Thomas (1994). *Using a Semantic Concordance for Sense Identificaction*. Proceedings of the ARPA Workshop on Human Language Technology.

Miller G., R. Beckwith, C. Fellbaum, D. Gross y K. Miller (1990). *Five papers on WordNet*. CSL Report 43, Cognitive Science Laboratory, Princenton University.

Miller, G. A., C. Leacock, R. Tengi y R. T. Bunker (1993). *A Semantic Concordance*, Proceedings of the ARPA Workshop on Human Language Technology.

Molina, A., F. Pla, L. Moreno y N. Prieto (1999). APOLN: A Partial Parser of Unrestricted Text. Proceedings of 5th Conference on Computational Lexicography and Text Research COMPLEX-99. pp 101-108. Pecs, Hungary.

Vossen P. (1996). *EuroWordnet: Building a multilingual wordnet database with semantic relations between words*. Technical and Financial Annex, EC funded project LE#4003.

Vossen P. (1998). *EuroWordnet: Building a Multilingual Database with word nets for European Languages*. In: K.Choukri, D.Fry, M.Nilson (eds.), the ELRA Newsletter, Vol3, n1.